Rejoinder

Andrew GELMAN, Gary KING, and Chuanhai LIU

We are grateful to the discussants for their thoughtful comments. We appreciate all of their remarks about our procedure in particular and multiple imputation for multiple surveys (MIMS) in general. We (and it seems our discussants) see the main contribution of our article as identifying an important statistical issue that had not before been completely recognized and providing one fairly general principled statistical approach to attack it. We recognize that many other specific models could be developed and encourage future researchers to pursue these potentially fertile research paths, several clearly laid out by the discussants. Put most generally, our main points are as follows:

1. A statistical method will be most useful when it can make use of all relevant available information (in our case using a multivariate imputation model, in the tradition of Fisher's (1932) method of covariance adjustment).
2. Any method rich enough to take advantage of all this information will by necessity be complex in interesting settings (with, in our case, a hierarchical model for between-survey variation).
3. In any application, diagnostic tools are necessary to check the resulting inferences and predictions.

These points apply to all areas that the discipline of statistics has reached, but for the problem we identify, the readily available ad hoc approaches are especially lacking, and so the opportunity for improving statistical practice may be unusually large.

We begin our rejoinder with a return to our motivating example of pre-election polling to illustrate our main contribution and to discuss the role of multiple imputations in a key analysis there. With this as background, we then discuss the several issues raised in the discussions, most of which address various assumptions and approximations in the imputation model. In an Appendix, we clarify the history of monotone data augmentation and describe a related recent improvement we made that sped up the MIMS algorithm by a factor of about five.

1. THE BENEFIT OF IMPUTATIONS IN OUR MOTIVATING EXAMPLE

To put yourself in our position at the start of this project, imagine what an applied social researcher might do in the setting of the example that motivated this work, in which we were interested in a real substantive question and willing to use the best statistical techniques available but not willing to invest the time to develop new ones.

We were confronted by the following problem in political science. At the time of the 1988 Presidential nominating conventions, national samples of survey respondents favored Michael Dukakis over George Bush by 17 percentage points. Yet on election day, Bush trounced Dukakis. We wanted to understand what had happened. From the political science literature (see Gelman and King 1993), it is known that the polls generally converge to the election outcome by election day, and so selection bias and measurement problems are not issues. The two Presidential campaigns are normally quite balanced, and so the result cannot be accounted for by differences in campaign skill, spending, verbal gaffes, spin doctors, or debating styles. It is also known that by the time of the election, citizen preferences normally become "enlightened" in that they depend more on variables fundamental to their interests such as characteristics of the candidates, economic conditions, ideological differences between voters and the two candidates, and demographics.

One can develop a new hypothesis (see Gelman and King 1993): what was unusual was not that Bush won, because that outcome was consistent with the values of the fundamental variables that year, but that Dukakis was so far in the lead (in the polls) early on. The reason for Dukakis's early lead, in this theory, was that voters' understanding of the two major party candidates was affected by the exceptionally long Presidential primary season, during which Dukakis eventually edged out Jesse Jackson for the Democratic Party's nomination but Bush ran uncontested (and hence essentially invisibly) for the Republican Party's nomination. During the primaries, Jackson's African-American heritage automatically made people think that Dukakis was more racially conservative (the politically more popular position in America) than even Bush, which was quite the reverse of reality. Consistent with this hypothesis is the fact that voters got it right by election day, accurately viewing Bush as more conservative.

To evaluate this theory, one more piece of evidence seems necessary. If the theory is right, then the predicted difference between black and white voter preferences, with other factors (e.g., party, ideology, region) held constant, should have been small at the start of the general election campaign, because of widespread confusion about the candidates' positions, but larger later on as people learned. To test this implication of the theory, one would like a set of cross-sectional surveys over the campaign; unfortunately, no single polling source was comprehensive enough to cover the necessary time span. (For example, CBS/*New York Times* provided an extensive series of surveys, but they did not adequately cover the Democratic and Republican nominating conventions.) The best available data are the 51 polls described in our article. We would like to run a logistic regression of voter preferences on ethnicity and the other control variables for each survey. The time trend in the predictive effect of ethnicity could then be examined to see how

Journal of the American Statistical Association
September 1998, Vol. 93, No. 443, Applications and Case Studies

differences between whites and blacks developed during the campaign.

Unfortunately, only five of the surveys asked all of the survey questions needed for our analysis. So, given existing methods, what can a researcher do? One might think of pleading with the survey organizations to include all questions in all surveys, but the campaign is already over (and the organizations surely included as many questions as they could afford). One could consider treating the missing questions as incomplete data and imputing them in each survey separately, but there is insufficient information for making any imputations. One might consider including all available variables in each of the 51 surveys, but then changes in the coefficients could be due to differing patterns of omitted variable bias. Another approach is to run the analysis just on the surveys with complete sets of variables, but this leaves only five time points, only one of which is within 4 months of election day.

We developed the MIMS approach to avoid the uncomfortable choice among these inadequate alternatives. Figure 1 compares the results of our multiple-imputation analysis to the leading ad hoc approach, a complete-case analysis including only those five surveys in which all of the questions were asked. Each graph plots the "first difference" of the predictive effect of ethnicity on Bush support (i.e., the difference in estimated probability of support for Bush between whites and blacks, holding all other variables constant at central scale values) by the date of the survey. Mean posterior estimates appear as solid circles, with error bars showing ± 1 within-imputation and total standard deviations. These first differences represent the change in probability of support for Bush, comparing whites to blacks with all other variables in the regression held constant. We would not interpret this as the causal effect of ethnicity on public opinion, because it is not clear how ethnicity would be defined as a causal "treatment" given that we are controlling for some of its consequences (see Rubin 1974b). Rather, we are interested in the time trend in this predictive quantity because it shows how preferences for blacks and whites differ, after controlling for the other political and demographic variables in the analysis. As an example, the first point on the multiple-imputation graph indicates that for our earliest poll, whites supported Bush only about 10% (plus or minus about 10%) more than blacks, holding constant other factors.

The key substantive issue is whether these first differences effects are increasing as hypothesized. The complete-case analysis [Fig. 1(b)] does not convey much information. So few points are available that the trend is not at all certain, but there is clearly no evidence that it is increasing. In fact, no information at all is provided during the last 5 weeks of the campaign, a critical period for this hypothesis. In contrast, the multiple-imputation analysis is far more informative. The regression line fit to the multiply imputed points is undoubtedly upwardly sloping (and significantly so), precisely as the hypothesis suggests. The difference between blacks and whites increases from 10% at the start of the campaign to about 30% by election day. Some of the multiply imputed points are more certain than others, but they all provide some information, and together they provide far more than the ad hoc complete-case approach.

For completeness, Figures 2 and 3 show similar first-difference graphs for each of the other variables in our logistic regressions from our multiple-imputation and complete-case analyses, with the variables listed in approximate order of predictive importance. In earlier work (Gelman and King 1993), with MIMS as yet unavailable, we worked around the missing-data problem by excluding the three variables that were the most missing—perceived ideological differences from Bush and Dukakis, and view of the economy—from the analysis. Unfortunately, these three variables are all highly predictive and also important to our substantive story. Perception of the economy is well known to be a crucial variable in deciding Presidential elections, and the ideological differences between respondents and candidates are particularly relevant given our discussion earlier about perceptions of Dukakis's ideology. Our logistic regression analyses tell a much clearer political story when they control for these variables.

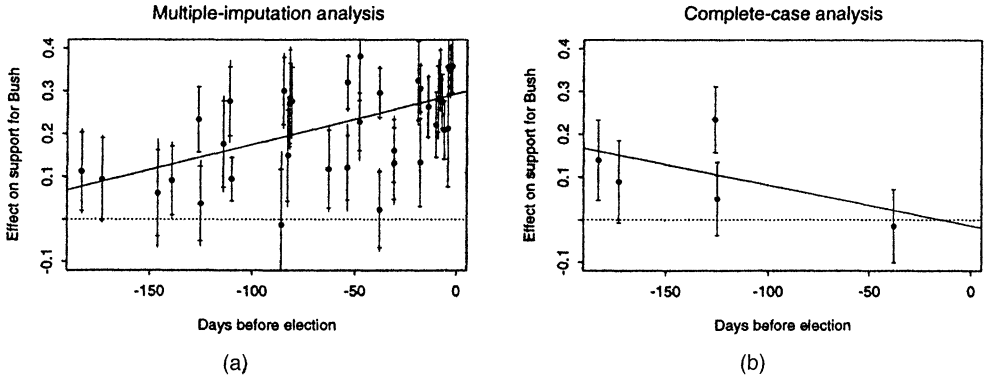

Figure 1. Comparison of Multiple-Imputation (a) and Complete-Case (b) Analyses for the Predictive Effect of Ethnicity on Support for Bush in the 1988 Polls. Only five polls asked enough questions to be included in the complete-case analysis, and only one of these polls was taken in the last 4 months of the election. Thus the complete-case analysis completely misses the trend that is apparent in the multiple-imputation analysis. See the captions for Figures 2 and 3 for explanation of the error bars.

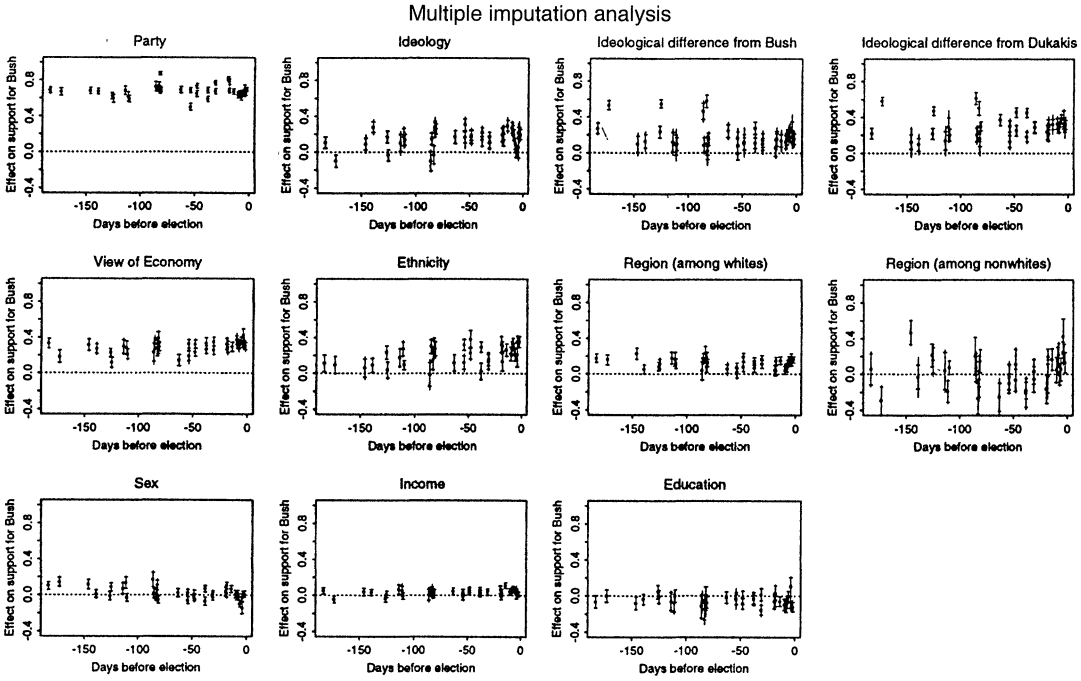

Figure 2. Predictive Effect Estimates for All of the Regression Variables in the Logistic Regression on Support for Bush in the 1988 Polls, Based on the Multiple-Imputation Analysis. Inner error bars show ±1 within-imputation standard errors (derived using the delta method from the logistic regression analysis); outer error bars show ±1 total standard errors (defined as the square root of the between-imputation variance plus the average within-imputation variance).

2. RESPONSES TO ISSUES RAISED IN THE DISCUSSIONS

2.1 Model Assumptions, Model Checking, and Model Improvement

Several discussants raise theoretical issues, including the frequency validity of multiple-imputation confidence statements (Binder, Judkins), the use of improper prior distributions (Kadane), and the accuracy of the multivariate model (Binder, Judkins) in our example and in general. These are all important theoretical points, and of course we cannot be completely confident about any of these issues in the context of a specific application, so our general approach to dealing with potential model violations is (1) to model as realistically as possible given the practical constraints imposed by our fitting procedure and (2) to check the fit of the model with respect to estimands of interest. With respect to (1), we try to include enough variables so that ignorability is at least a reasonable (if not perfect) assumption, and set up the model with some care (as, for example, in transforming the age variable into categories). With respect to (2), we believe that our proposed methods of model checking are crucial, and we would not trust the multiple-imputation analysis in any specific application without seeing some diagnostic results.

In general, a strength of any model-based procedure is that it is based on assumptions that can be checked and, if found to be flawed, improved. The discussants suggest several places in which the imputation model has clear potential for improvement, including modeling discrete responses (Kadane), using informative prior distributions on time-series patterns (Santos), and including more variables and interactions (Binder). All of these suggestions make sense, and a natural extension of our model is to have latent multivariate normal distributions with ordered multinomial logit or probits for the discrete observations. One could then use this model to impute both the "underlying" continuous responses and the discrete missing observations. In the same spirit, the method might be improved by adding more variables to the continuous model, replacing the linear trend with more complicated time-series patterns for the mean function, letting the variance matrix vary over the surveys, and, in our application, simultaneously modeling voter preferences with likely voter turnout.

We have not made these improvements yet, because we believe that our current procedure (modeling combined with diagnostics) can be extremely useful compared to existing alternatives. Improvements in the model should of course make the methods even more useful. Multivariate modeling of missing data is an open research area, and we expect that as multiple imputation is applied in more and more applied settings, we will learn more about what works well in practice. Of course, as several discussants note, no imputed dataset is as good as convincing the survey organizations to ask all relevant questions and designing the survey to minimize unit and item nonresponse.

2.2 Ignorability Assumptions

All of the discussants agree, as do we, that it can be risky to assume ignorable nonresponse. This is of course a problem with all survey analyses, whether or not imputation is used. Complete-case and available-case methods are far more sensitive to ignorability assumptions because they generally are reliable only if data are missing com-

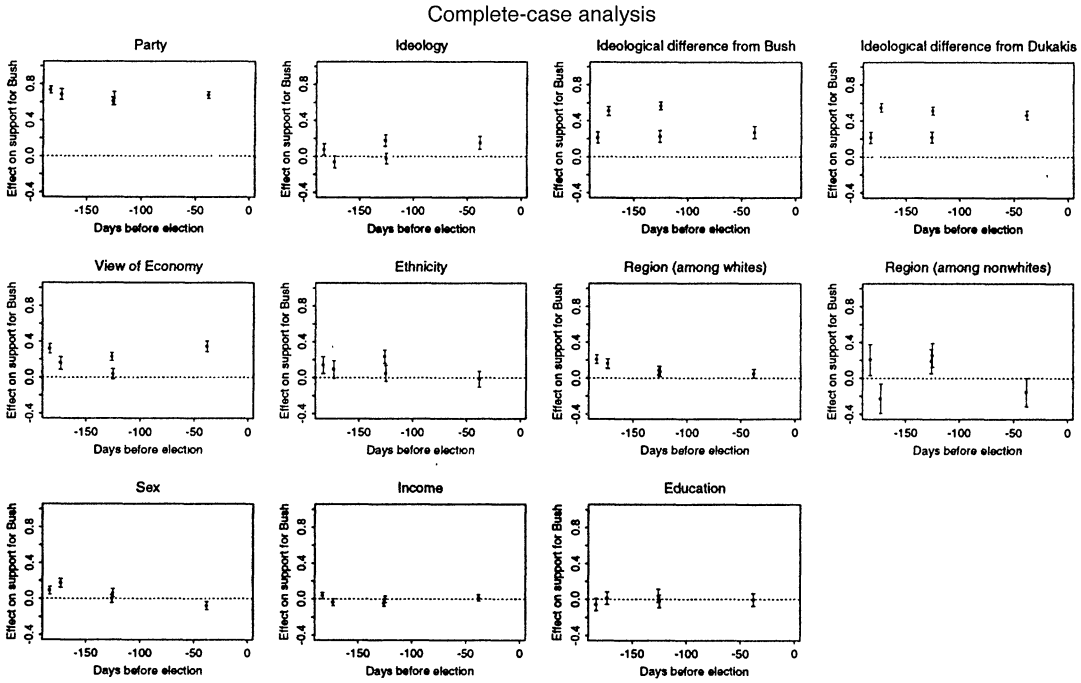

Figure 3. Predictive Effect Estimates for All of the Regression Variables in the Logistic Regression on Support for Bush in the 1988 Polls, Based on the Complete-Case Analysis. Error bars show ±1 standard errors derived using the delta method from the logistic regression analysis. Only five surveys were available in which all questions were asked, although every question was asked in at least 10 surveys (see Table 1 of the main article).

pletely at random (MCAR), whereas multivariate imputation methods assume the weaker condition of missing at random (MAR). The key generalization in our procedure, relative to other available missing-data imputation methods, is to unasked questions in multiple surveys. And the pattern of nonresponse due to unasked questions is quite reasonably approximated by an ignorable model, because the entire population is "missing" in these cases. (In fact, all of the discussants' issues regarding ignorability are for within-survey nonresponse, not for our main goal of coping with unasked questions.)

2.3 Use of Design Information

Some of the discussants commented on our use of design information by computing weighted averages to estimate survey means. Our goal here was not to model the design so much as to include the knowledge used by the survey organizations. As discussed by Voss, Gelman, and King (1995), most of the surveys that we analyzed use very similar designs (simple random samples of households with some geographic stratification, and then some selection of an adult within sampled households). However, they varied greatly in the information included in the survey weights. Almost all of the surveys weighted on sex, many also adjusted for other demographic variables, and some also adjusted for telephones and household size. (Judkins states that "poststratification is usually a minor adjustment" compared to weighting for household size and number of telephone lines, but in many of our surveys, poststratification was the only weighting used. Even in the surveys that included adjustments for household size and telephones, the poststratification weights were not minor. For example, in one of the CBS polls, a white female age 18–29 from the West with a college degree would have a poststratification weight of .30, whereas a black male age 18–29 from the South without a high school degree would have a poststratification weight of 2.25.)

In our analysis, we respect the judgment of the survey organizations that the information used in the weighting for each survey is necessary for the surveys to match the population. We do not try to improve on the survey weightings (although we believe that this is possible; see Gelman and Little 1998); rather, we try to approximate the analysis that would occur if all of the questions were asked in all the surveys, but with the existing weighting schemes. This approach is "design consistent" in the practical sense of accounting for the design factors judged to be important by the original survey organizations. This approach is also consistent with Brehm's point that many "survey responses" are not even defined until the question is asked. Rather than trying to model any underlying response, the multiple-imputation procedure merely models the responses that would be observed if the questions were asked and answered for all respondents (which is the most that any imputation procedure can reasonably be expected to do).

2.4 Imputation versus Direct Modeling

Our procedure follows a standard multiple imputation prescription (Rubin 1987): fit a multivariate model to all the data, use it to impute the missing data, and then discard the imputation model and work from the completed datasets. Kadane makes a suggestion that is quite natural from the Bayesian point of view: Why not skip the imputation stage entirely and directly fit a realistic model to the

observed data? Our short answer is that we believe that our current model is good enough for imputing the relatively small fraction of missing data, but not good enough to replace all of the observed data. We believe this partly because, before our current program was available, we (along with most other social scientists) were dealing with missing data using a patchwork of complete-case, available-case, regression, and imputation models.

More generally, there is a theoretical basis for our unwillingness to jump to the full Bayesian modeling approach in this case. To perform our imputation, we need a multivariate model, which by necessity requires many parameters and would require an immense research effort to be realistic, given all of the potential interactions between variables. If we had no missing data, however, we would base much of our substantive analysis on time series plots of cross-tabulations (as in the article's Fig. 5) and regression models (as in the figures in this rejoinder). Such models are relatively simple because they are conditional. Creating a realistic model of $p(y|X)$ is a much more modest task than realistically modeling $p(y, X)$, especially if X has many dimensions. The logic is that we would be happy with a good model for $p(y|X)$, but we are missing X for some units, so we impute using a reasonable approximate model for $p(y, X)$. This can work well if not too much of X is missing. Meng (1994) has provided more discussion of the effects of using different models for imputation and analysis.

2.5 Unasked Questions, Split Forms, Panel Surveys, and Other Applications

One attractive area of application is split forms, matrix sampling, and other settings in which different respondents are asked different questions. Although, as Judkins notes, "there are few surveys where the set of questions varies across strata," such designs can in fact be useful in a surveys with large numbers of potential questions (see Raghunathan and Grizzle 1995). Other natural applications of such designs include political surveys in which different questions can be asked to respondents at different times (e.g., before and after a debate) or in different geographic areas. In addition, with the growing use of computer-aided interviewing, it is increasingly common for different respondents to be asked different questions, whether as part of an experiment (as suggested by Brehm) or simply to streamline the interviewing process.

Santos points out that panel surveys are a natural application for our methodology. In this setting, in addition to modeling between-survey effects, one could fit a hierarchical model to random effects for panel respondents as well.

As an aside, to fit a hierarchical multiple imputation model it is not necessary that there be a stratum in which all the questions are asked; as noted by Judkins, it is only required that every *pair* of questions be asked together, which gives enough information to estimate the pairwise correlations that characterize a multivariate normal model.

2.6 General Comments on Missing-Data Imputation

We believe that future researchers analyzing multiple surveys with unasked and unanswered questions will agree with our main argument and see the value of imputation approaches that make use of all available information (as in our multivariate imputation model) while accounting for within- and between-survey variation (as in our hierarchical model). However, we can see many reasons to modify, extend, or even approximate the specific statistical imputation model we developed for our application. As Binder notes (also see Fay 1996; Rubin 1996), a variety of different methods are used to impute missing data in practice. We would like our work to be useful not just to users of our program, but also to missing-data imputers in general. In that spirit, what lessons have we learned that can be applied to missing-data imputation in general?

Because of the complexity of the required multivariate models, checking the fit of multiple-imputation models is essential. This can be done, as we did, by running internal checks by cross-validation and by comparing the imputed data to those observed. Our specific diagnostics (as illustrated in the figures in the article) will probably be of use to most, but some form of checking should always be conducted.

Using multiple rather than single imputations allows us to capture imputation uncertainty, which can then be included in all graphical displays, as, for example, in the within- and between-imputation error bars in our figures. (In a related issue, Judkins notes that that hierarchical Bayes estimates tend to be drawn toward the regression line and as a result tend to be insufficiently variable. This problem is resolved with multiple imputation, however; Judkins's claim that "the random effects are assumed to be zero for surveys where the corresponding questions were not asked" is in fact not true for our procedure, in which the random effects are drawn from multivariate normal distributions.) Imputing posterior means rather than simulation draws would lead to many serious problems, including attenuation of estimated regression coefficients.

Finally, using some type of multivariate model seems essential so that the observed parts of an outcome variable can be used with other information to impute the missing parts of the explanatory variables. Although it seems counterintuitive to many who want to impute only in some causal order, using y to impute X induces no endogeneity bias if draws are from the full posterior distribution. A related source of confusion is the random imputation variance added to the imputed values in the explanatory variables. Because this variability is drawn from the posterior distribution, it does not attenuate the regression coefficients as random measurement error can.

We know ahead of time that any imputation method will be imperfect, and we must explore the data and imputations to find these imperfections and understand their practical implications. But these potential problems should not lead us to rely on more seriously flawed methods that try to sidestep the missing-data issue.

APPENDIX: COMPUTATIONALLY EFFICIENT IMPUTATIONS

Here we respond to Kadane's inquiry about the history of monotone data augmentation (MDA) and add our latest improvement to the MIMS algorithm. Maximum likelihood estimation of a multivariate normal distribution from data with monotone patterns of missingness dates to work of Anderson (1957). The technique used is commonly known as the "factored likelihood approach," because it makes use of the fact that the multivariate normal distribution allows for a likelihood factorization with distinct parameters (Rubin 1974a) even with monotone ignorable missing data. The use of the factored likelihood approach, or analysis of variance for regression, in the context of the decomposition of the Wishart distribution dates back to work of Bartlett (1933, 1939). For generating Wishart matrices, Odell and Feiveson (1966) developed a computationally more efficient algorithm than the factored likelihood approach used by Hartley and Harris (1963). Odell and Feiveson's algorithm is based on the Bartlett decomposition of the Wishart distribution, which can be stated informally as follows: The Wishart matrix can be written as a product of a triangular matrix and its transpose, where the elements of the triangular matrix are independent χ and normal variables (see, e.g., Anderson 1984, pp. 249–251, Kshiragar 1959, and the references therein).

In the context of multiple imputation for single surveys, Rubin and Schafer (1990) proposed the MDA algorithm and implemented MDA using the factored likelihood approach. As Kadane notes, the relevant theoretical results have appeared in various places, including the work of Chen (1986). Extending the results of the Bartlett decomposition, Liu (1993) provided a computationally more efficient algorithm for implementing MDA than that based on the factored likelihood approach.

We began our research with the specific MDA algorithm of Rubin and Schafer (1990). We modified and implemented MDA using the results given in our Theorems 1 and 2, because they allow for a more efficient computation than that implemented using the factored likelihood approach. With our revised MDA algorithm, a monotone pattern is created based on all of the observed data from all of the surveys. The constructed monotone pattern can contain structured missing patterns resulting from the fact that some of the questions were not asked in some of the surveys and some of the questions asked were not answered. Recently, we have implemented an improved version of our algorithm in which individuals are ordered according to their missing data patterns so that CPU time is reduced in applying the sweep operator when imputing the missing data that destroy the monotone pattern (step 1). As a result, our current implementation of the algorithm runs about five times faster than our old implementation without grouping.

ADDITIONAL REFERENCES

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Bartlett, M. S. (1933), "On the Theory of Statistical Regression," *Proceedings of the Royal Society of Edinburgh*, 53, 260–283.

——— (1939), "A Note on Tests of Significance in Multivariate Analysis," *Proceedings of the Cambridge Philosophical Society*, 35, 180–185.

Fisher, R. A. (1932), *Statistical Methods for Research Workers* (4th ed.), Edinburgh: Oliver and Boyd.

Gelman, A., and Little, T. C. (in press), "Improving Upon Probability Weighting for Household Size," *Public Opinion Quarterly*.

Hartley, H. O., and Harris, D. L. (1963), "Monte Carlo Computation in Normal Correlation Problems," *Journal of Association for Computing Machinery*, 10, 301–306.

Kshiragar, A. M. (1959), "Bartlett Decomposition and Wishart Distribution," *Annals of Mathematical Statistics*, 30, 239–241.

Odell, P. L., and Feiveson, A. H. (1966), "A Numerical Procedure to Generate a Sample Covariance Matrix," *Journal of the American Statistical Association*, 61, 199–203.

Rubin, D. B. (1974a), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1974b), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.